An Evaluation of Graded Sense Disambiguation using Word Sense Induction

David Jurgens
HRL Laboratories, LLC *and*
University of California, Los Angeles

Talk overview

- Word Sense Disambiguation and Graded Senses Assignment
- Word Sense Induction
- Evaluating Graded Sense Assignments

Talk overview

- Word Sense Disambiguation and Graded Senses Assignment
- Word Sense Induction
- Evaluating Graded Sense Assignments

Word Sense Disambiguation

Goal: Given a usage of a word, determine which meaning is present

John sat on the chair.

1. a seat for one person, with a support for the back
2. the position of professor
3. the officer who presides at the meetings of an organization

Polysemy makes sense disambiguation much harder.

Mary handed the paper to her professor.

1. a material made of cellulose pulp
2. an essay, especially one written as an assignment
3. a daily or weekly publication on folded sheets
4. a medium for written communication
5. a scholarly article describing the results of observations
6. a business firm that publishes newspapers

Polysemy makes sense disambiguation much harder.

Mary handed the paper to her professor.

1. a material made of cellulose pulp
2. an essay, especially one written as an assignment
3. a daily or weekly publication on folded sheets
4. a medium for written communication
5. a scholarly article describing the results of observations
6. a business firm that publishes newspapers

Polysemy makes sense disambiguation much harder.

After working all night writing her assignment,
Mary handed the paper to her professor.

1. a material made of cellulose pulp
2. an essay, especially one written as an assignment
3. a daily or weekly publication on folded sheets
4. a medium for written communication
5. a scholarly article describing the results of observations
6. a business firm that publishes newspapers

Polysemy makes sense disambiguation much harder.

After working all night writing her assignment,
Mary handed the paper to her professor.

1. a material made of cellulose pulp
2. an essay, especially one written as an assignment
3. a daily or weekly publication on folded sheets
4. a medium for written communication
5. a scholarly article describing the results of observations
6. a business firm that publishes newspapers

Most applicable sense

Moderately applicable sense

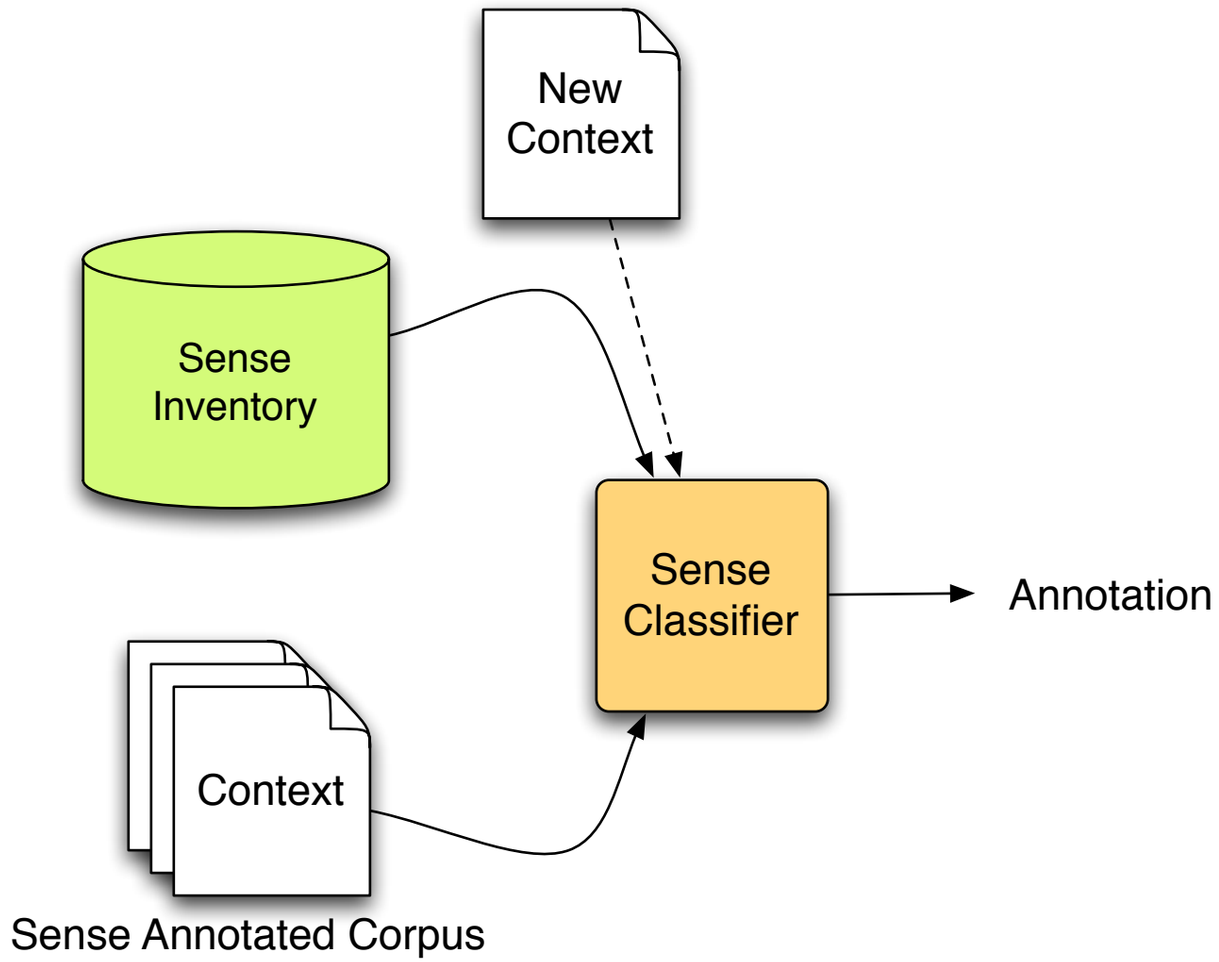
The goal of Graded Word Sense Disambiguation

- Identify which senses are salient in a given context
- Quality the senses' degree of applicability

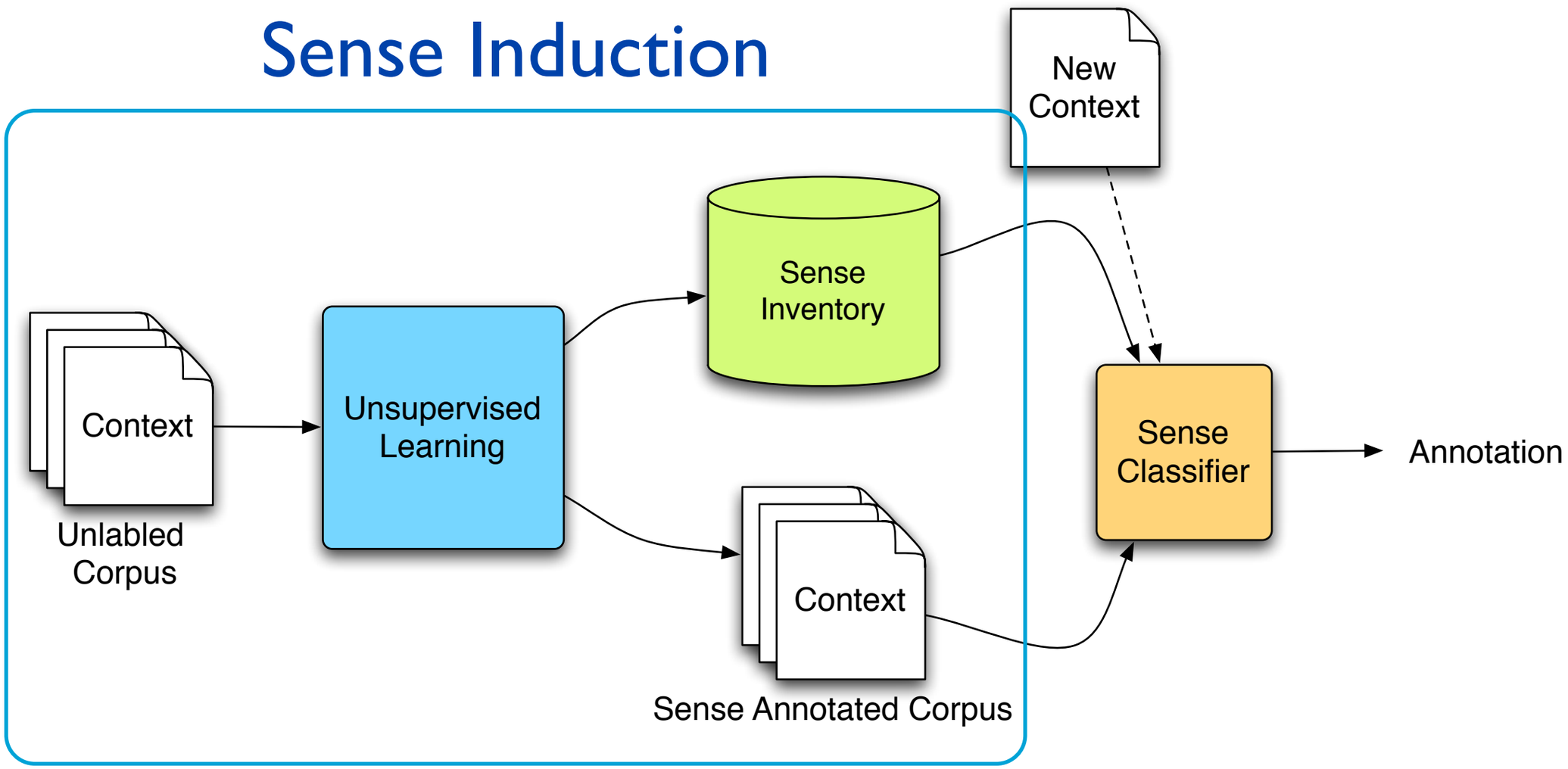
Talk overview

- Word Sense Disambiguation and Graded Senses Assignment
- Word Sense Induction
- Evaluating Graded Sense Assignments

Word Sense Disambiguation



Word Sense Induction automatically learns the sense inventory from contextual examples



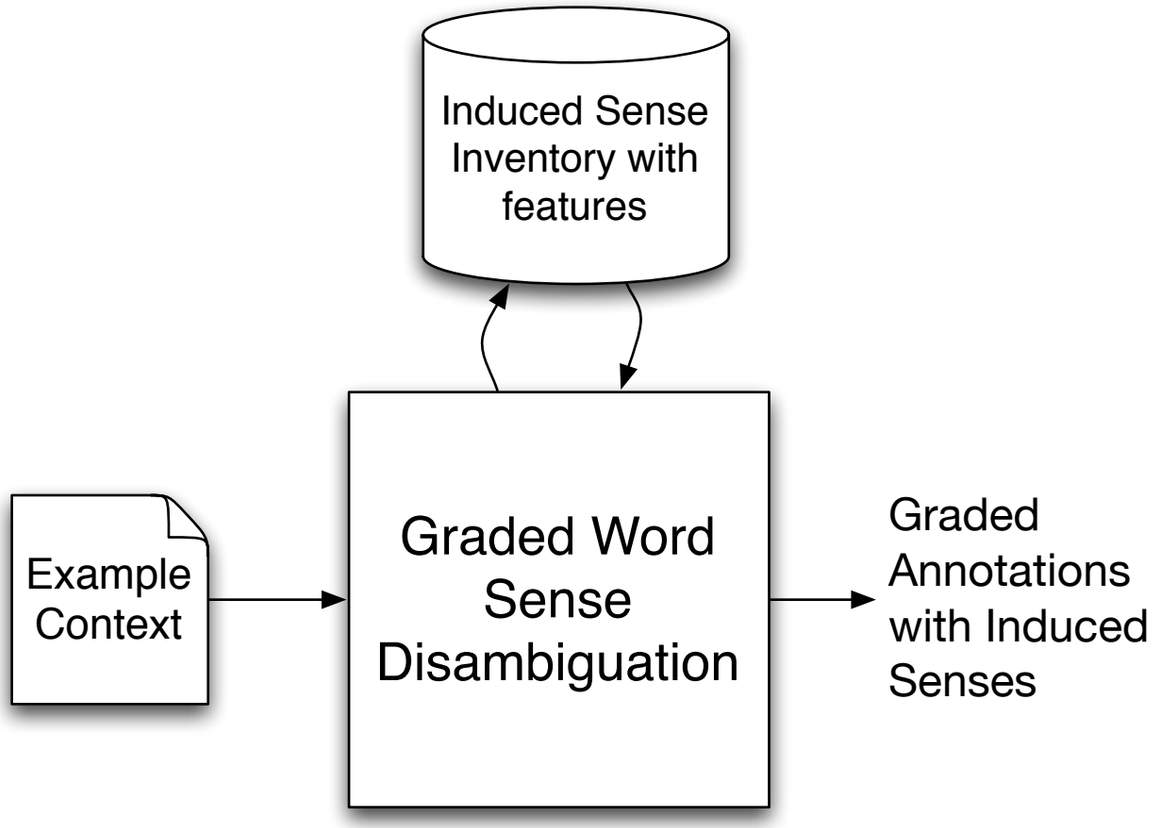
Benefits of Sense Induction

- Discovery of corpus-specific senses
- Removes the need for learning sense features from sense-annotated corpora
- No more annotation bottleneck!

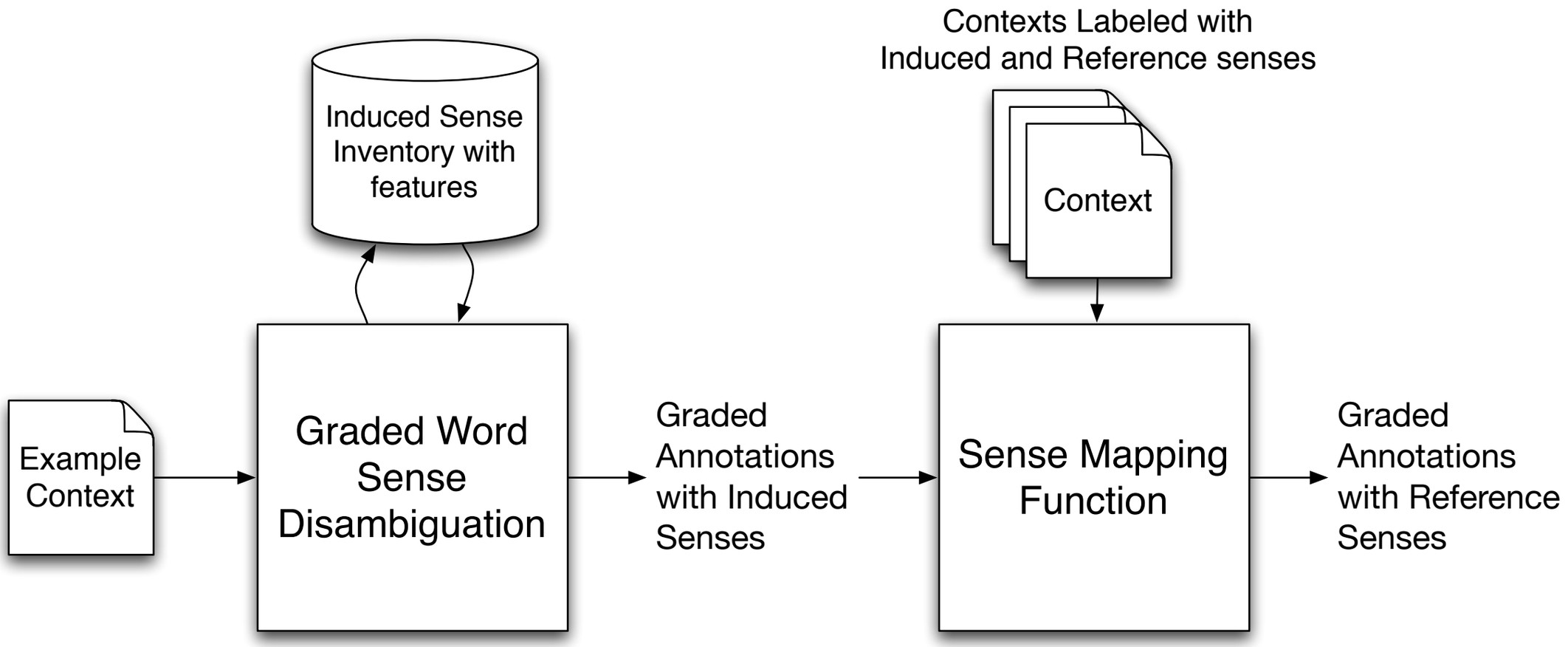
Potential Drawbacks

- What senses are being learned?
- How useful are the learned senses?

Induced senses can be used directly, but often a specific sense inventory is needed



Induced senses can be used directly, but often a specific sense inventory is needed



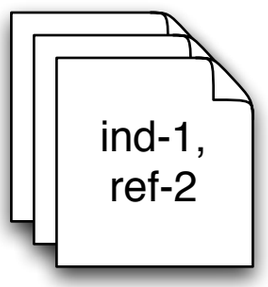
Produces an end-to-end WSD system with the desired sense labels, using minimal labeled data

Three graph-based WSI methods tested

	Korkontzelos and Manandhar (2010)	Jurgens (2011)	Navigli and Crisafulli (2010)
Sense Induction Scope	Nouns, Collocated Nouns	Nouns, Verbs, and Adjectives	Nouns, Verbs, and Adjectives
Graph Features	Single Word	All Words	Single Word
Sense Discovery Method	Chinese Whispers (Biemann, 2006)	Link Clustering (Ahn, 2010)	Edge Deletion

Sense Remapping at a high level

1. Co-label corpus with both sense types



2. Build a collocation matrix for senses

	r_1	r_2	r_3
i_1	1	5	
i_2			6
i_3	8		

3. Construct a function to remap specific instances

$$i_i \rightarrow \{r_j\}$$

Naïve approach: map each induced sense to a single sense

	r_1	r_2	r_3
i_1	6	12	2
i_2	10	2	4
i_3	2	16	2

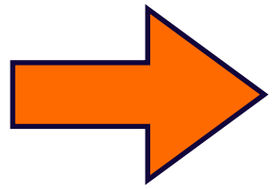
Step 1: Count co-labeling of induced and reference senses

	r_1	r_2	r_3
i_1	0.3	0.6	0.1
i_2	0.625	0.125	0.25
i_3	0.1	0.8	0.1

Step 2: normalize row values to be $p(r_i|i_j)$.

Naïve approach: map each induced sense to a single sense

	r_1	r_2	r_3
i_1	0.3	0.6	0.1
i_2	0.625	0.125	0.25
i_3	0.1	0.8	0.1



$i_1 \rightarrow r_2$

$i_2 \rightarrow r_1$

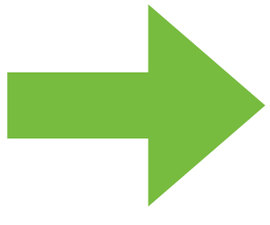
$i_3 \rightarrow r_2$

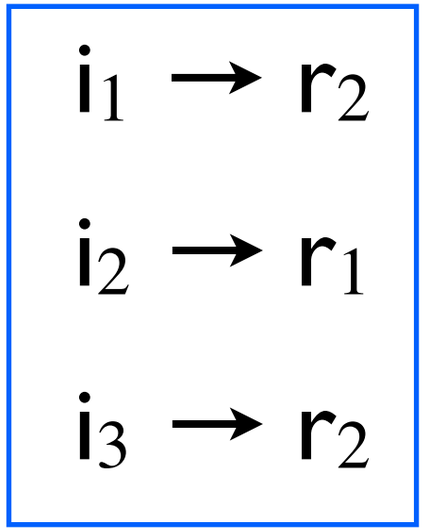
Step 3: Map each induced sense to the most likely reference sense

Naïve approach: map each induced sense to a single sense

Context with Graded Sense Labels

i_1	i_2	i_3
0.7	0.2	0.1





Sum the ratings of each of the remapped senses

r_1	r_2	r_3
0.2	0.8	0

Sense mapping from Agirre et al. (2006)

Step 3: multiply graded labeling by mapping matrix

i_1	i_2	i_3
0.3	0.5	0.2

*

	r_1	r_2	r_3
i_1	0.3	0.6	0.1
i_2	0.625	0.125	0.25
i_3	0.1	0.1	0.8

=

r_1	r_2	r_3
0.423	0.263	0.315

Select the highest weighted sense

Graded Variant: Relabel with *all* senses

Step 3: multiply graded labeling by mapping matrix

i_1	i_2	i_3
0.3	0.5	0.2

*

	r_1	r_2	r_3
i_1	0.3	0.6	0.1
i_2	0.625	0.125	0.25
i_3	0.1	0.1	0.8

=

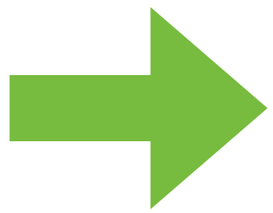
r_1	r_2	r_3
0.423	0.263	0.315

Use all the senses

Issue: the matrix construction method ignores the graded ratings

Contexts

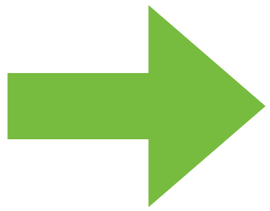
c_1	$i_1 : 0.9$ $i_2 : 0.1$	$r_1 : 0.9$ $r_2 : 0.1$
c_2	$i_1 : 0.1$ $i_2 : 0.9$	$r_1 : 0.1$ $r_2 : 0.9$
c_3	$i_1 : 0.1$ $i_2 : 0.9$	$r_1 : 0.1$ $r_2 : 0.9$



	r_1	r_2
i_1	3	3
i_2	3	3

Alternative: construct the co-label matrix by summing the product of the sense ratings

Contexts		
c_1	$i_1 : 0.9$ $i_2 : 0.1$	$r_1 : 0.9$ $r_2 : 0.1$
c_2	$i_1 : 0.1$ $i_2 : 0.9$	$r_1 : 0.1$ $r_2 : 0.9$
c_3	$i_1 : 0.1$ $i_2 : 0.9$	$r_1 : 0.1$ $r_2 : 0.9$



	r_1	r_2
i_1	0.83	0.27
i_2	0.27	0.83

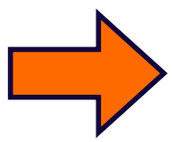
Compute the subsequent steps the same as Agirre et al. (2006)

Equivalent to the Naïve approach with a vector mapping

Steps 1 and **2**: construct row-normalized matrix as before

Step 3: Map each induced sense to a distribution over senses

	r_1	r_2	r_3
i_1	0.3	0.6	0.1
i_2	0.625	0.125	0.25
i_3	0.1	0.8	0.1



$i_1 \rightarrow$

r_1	r_2	r_3
0.3	0.6	0.1

$i_2 \rightarrow$

r_1	r_2	r_3
0.625	0.125	0.25

$i_3 \rightarrow$

r_1	r_2	r_3
0.1	0.8	0.1

Equivalent to the Naïve approach with a vector mapping

Context with Graded Sense Labels

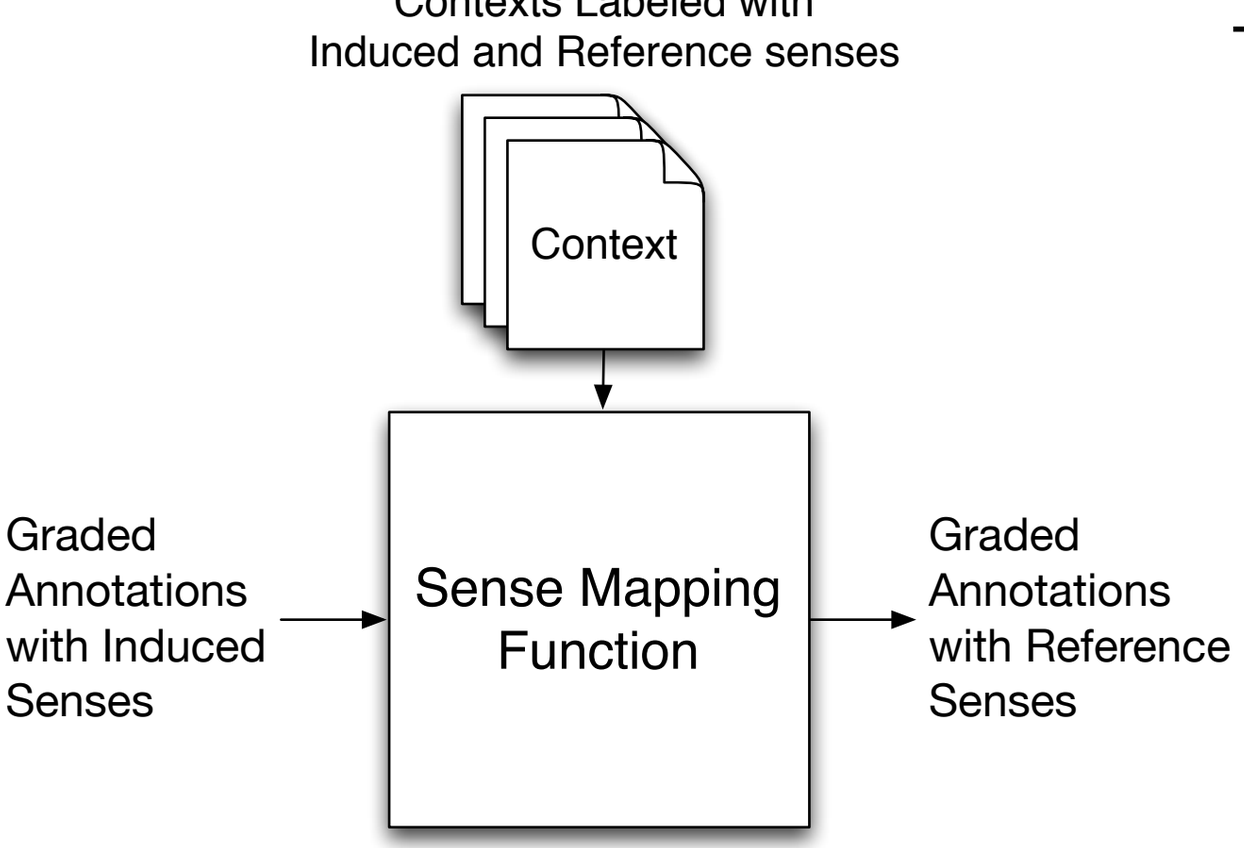
i_1	i_2	i_3
0.7	0.2	0.1

	r_1	r_2	r_3
$i_1 \rightarrow$	0.3	0.6	0.1
$i_2 \rightarrow$	0.625	0.125	0.25
$i_3 \rightarrow$	0.1	0.8	0.1

Sum the induced senses' distributions, weighted by their ratings

r_1	r_2	r_3
0.345	0.525	0.13

Remapping Summary



Three Approaches

- Naïve
- Agirre et al. (2006)
  - Vector
  - Vector + Matrix reweighting

Talk overview

- Word Sense Disambiguation and Graded Senses Assignment
- Word Sense Induction
- Evaluating Graded Sense Assignments

Evaluating Graded Sense Annotation

She *won* the gold medal with her hard work.

- wn_1: 0.6
- wn_2: 0.4
- wn_3: 0.0
- wn_4: 0.0

Multiple evaluation objectives

She *won* the gold medal with her hard work.

wn_1: 0.6
wn_2: 0.4
wn_3: 0.0
wn_4: 0.0

1. Which senses are applicable?

Multiple evaluation objectives

She *won* the gold medal with her hard work.

wn_1: 0.6

wn_2: 0.4

wn_3: 0.0

wn_4: 0.0

1. Which senses are applicable?
2. How do senses differ in their applicability?

Multiple evaluation objectives

She *won* the gold medal with her hard work.

wn_1: 0.6
wn_2: 0.4
wn_3: 0.0
wn_4: 0.0

1. Which senses are applicable?
2. How do senses differ in their applicability?
3. What is the applicability of each sense?

Detection: which senses are present?

She *won* the gold medal with her hard work.

Gold	Test$_1$	Test$_2$	Test$_3$
wn$_1$: 0.6	wn$_1$: 0.7	wn$_1$: 0.0	wn$_1$: 0.3
wn$_2$: 0.4	wn$_2$: 0.3	wn$_2$: 1.0	wn$_2$: 0.0
wn$_3$: 0.0	wn$_3$: 0.0	wn$_3$: 0.0	wn$_3$: 0.5
wn$_4$: 0.0	wn$_4$: 0.0	wn$_4$: 0.0	wn$_4$: 0.2
Detection:	1.0	0.5	0.25

Jaccard Index

$$\frac{G \cap T}{G \cup T}$$

Ranking: order the senses by their applicability

She *won* the gold medal with her hard work.

Gold	Test$_1$	Test$_2$	Test$_3$
wn$_1$: 0.6	wn$_1$: 0.7	wn$_1$: 0.0	wn$_1$: 0.3
wn$_2$: 0.4	wn$_2$: 0.3	wn$_2$: 1.0	wn$_2$: 0.0
wn$_3$: 0.0	wn$_3$: 0.0	wn$_3$: 0.0	wn$_3$: 0.5
wn$_4$: 0.0	wn$_4$: 0.0	wn$_4$: 0.0	wn$_4$: 0.2
Ranking:	1.0	0.33	-0.2

Goodman and Kruskal's γ

- Computed much like Kendall's τ
- Most appropriate for comparing rankings with many ties

Perception: quantify the applicability of each sense

She *won* the gold medal with her hard work.

Gold	Test$_1$	Test$_2$	Test$_3$
wn$_1$: 0.6	wn$_1$: 0.7	wn$_1$: 0.0	wn$_1$: 0.3
wn$_2$: 0.4	wn$_2$: 0.3	wn$_2$: 1.0	wn$_2$: 0.0
wn$_3$: 0.0	wn$_3$: 0.0	wn$_3$: 0.0	wn$_3$: 0.5
wn$_4$: 0.0	wn$_4$: 0.0	wn$_4$: 0.0	wn$_4$: 0.2
Perception:	0.98	0.55	0.41

Cosine Similarity

$$\frac{G \cdot T}{||G|| * ||T||}$$

Baselines

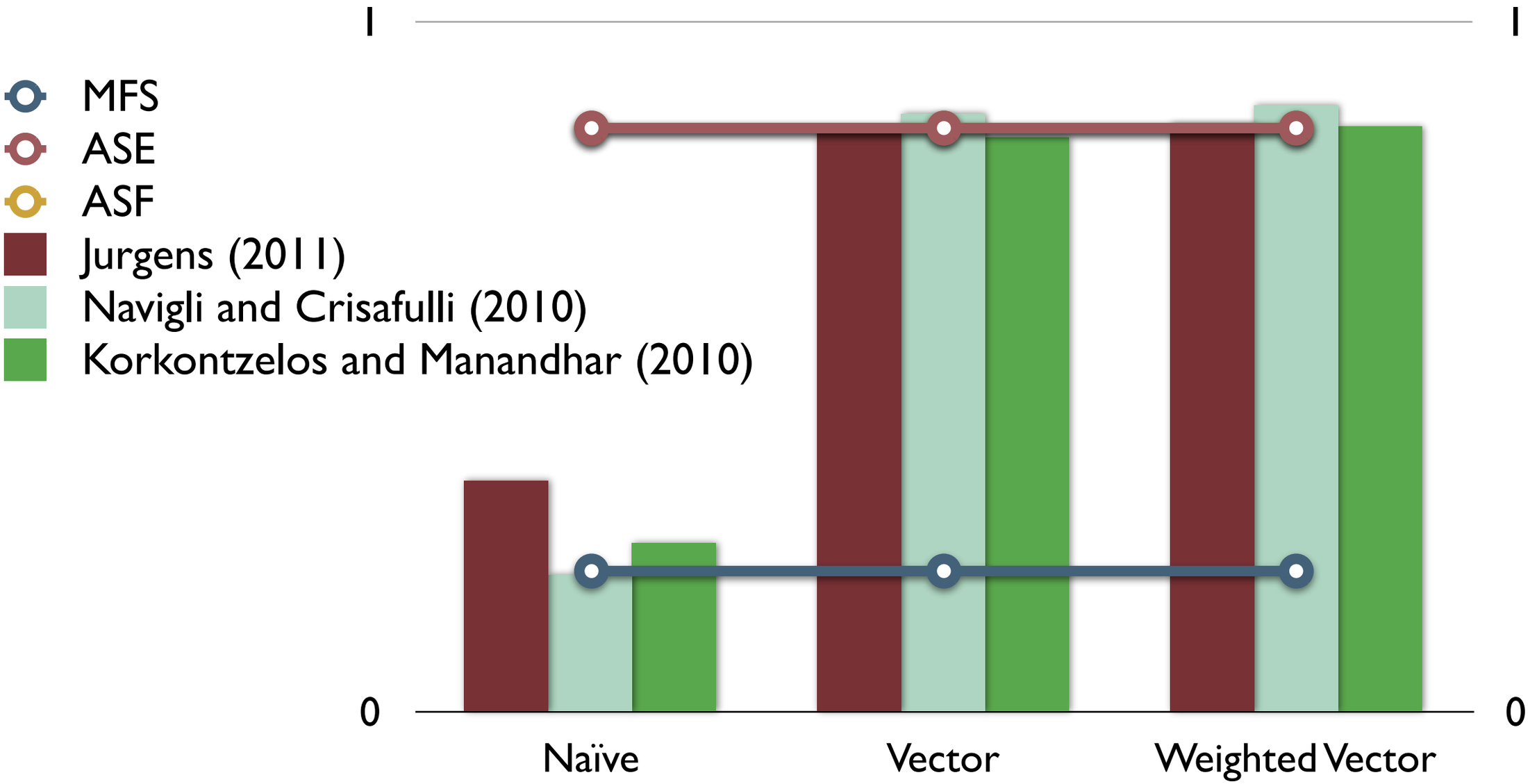
Sense assignments have a strong bias towards more frequent senses

- Most Frequent Sense (MFS)
- All Senses, ranked by Frequency (ASF)
- All Senses, equally ranked (ASE)

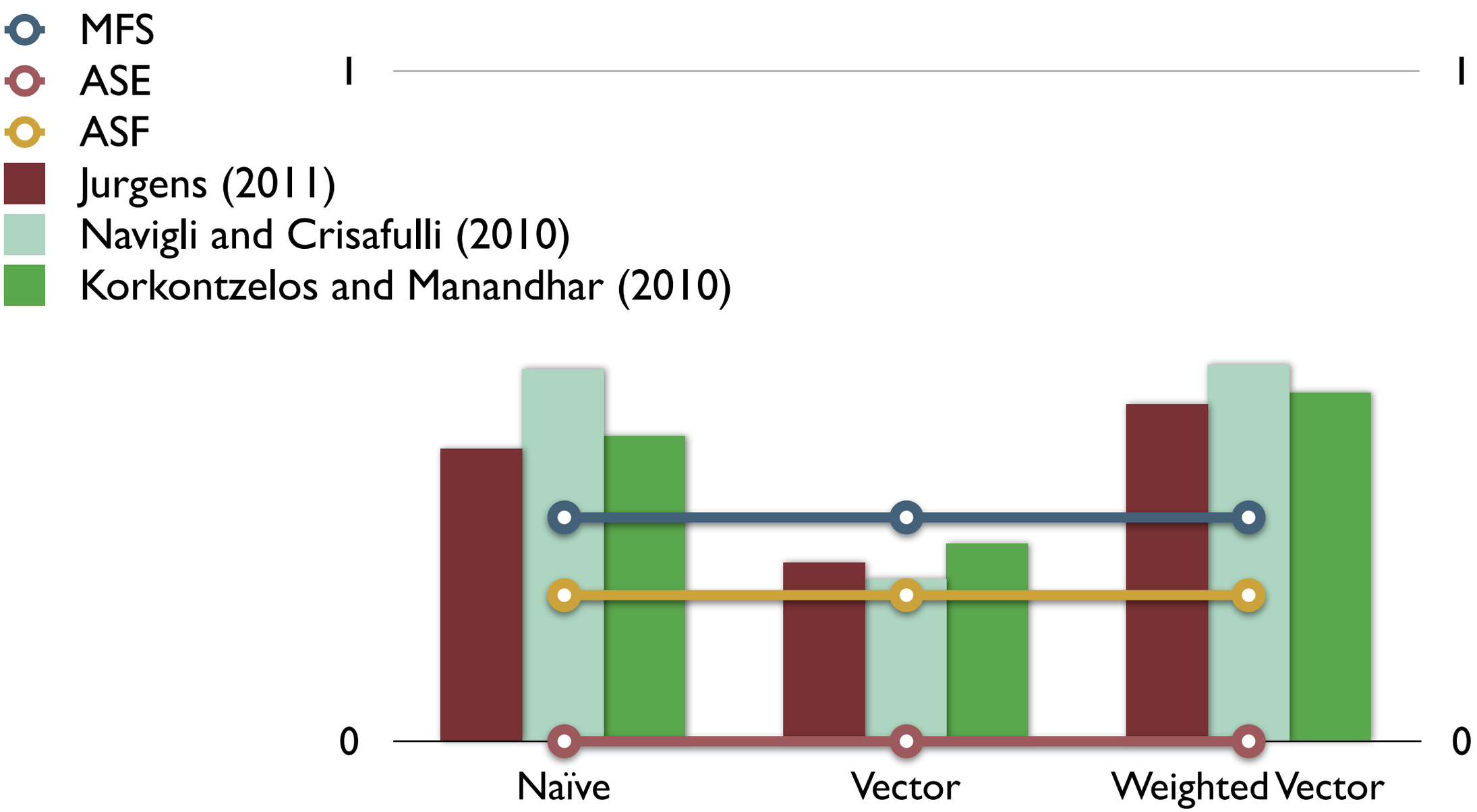
Experimental Setup

- Trained using the GWS Corpus provided by Erk et al. (2009)
- 8 terms: 3 nouns, 3 verbs, 2 adj. (4-7 senses each)
  - 50 contexts each
  - 3 annotators per context, average sense scores used as gold standard
- Used 80% to build sense mapping, 20% to test with cross validation

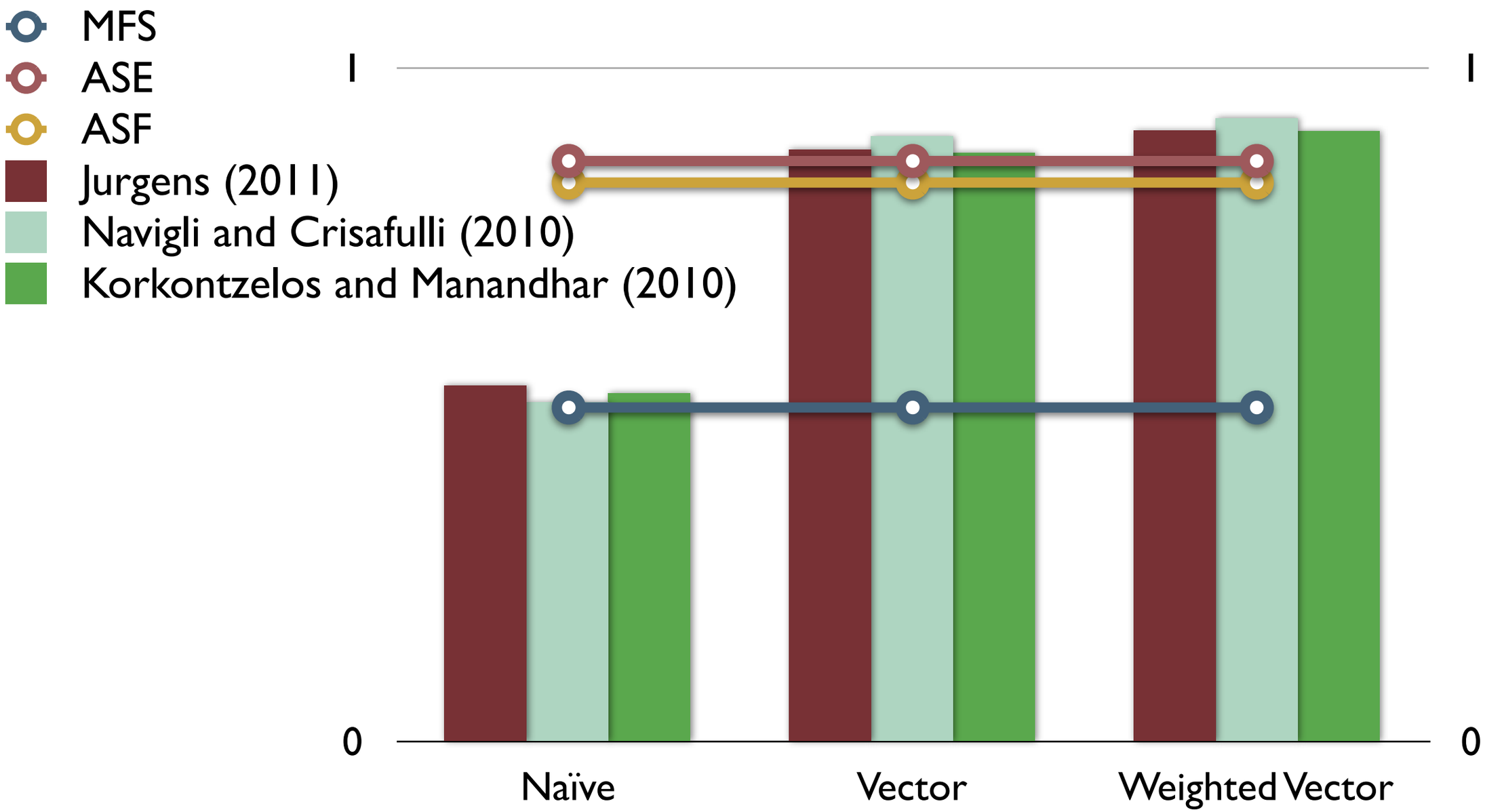
Perception Results



Ranking Results



Perception Results



Further Observations

- Approaches varied wildly in the number of senses labeled
  - Navigli and Crisafulli (2010) only labeled ~56%, but had very high precision (only 5% had > 1 label)
  - Jurgens (2011) ~2 senses on average to four words, and 60+ for the other four
- Ranking is much harder than Perception
  - A > B vs. A=.51 and B=.49

Take-aways

- WSI systems offer significant promise for automating Graded WSD
  - Seeding corpus annotation when WSD systems have too little training data
- Accounting for graded rating is essential in sense-mapping

Future Work

- Rethinking the *Perception* evaluation metric
  - Neither Cosine similarity nor Jensen-Shannon Diverge appear to be ideal measures
- Intrinsic clustering evaluations for partial cluster membership
- Task-based Evaluation for graded senses

Thank you!

jurgens@cs.ucla.edu

All models and data released as
a part of the S-Space Package

https://github.com/fozziethebeat/S-Space

Thanks to Katrin Erk, Diana McCarthy, and Nicholas Gaylord
for making GWS corpus available